DOCUMENT RESUME

ED 322 202 TM 015 407

AUTHOR Becker, Betsy Jane; Hedges, Larry V.
TITLE Synthesizing Correlation Matrices.
PUB DATE Apr 90
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; *Equations (Mathematics); *Estimation (Mathematics); *Mathematical Models; *Matrices; Regression (Statistics); *Synthesis
IDENTIFIERS *Correlation Matrices; Linear Models

ABSTRACT

The problem of combining information to estimate standardized partial regression coefficients in a linear model is considered. A combined estimate obtained from the pooled correlation matrix is proposed, and its large sample distribution is obtained. This estimate can be generalized to address situations in which not every study measures every variable. The method can be used to synthesize estimates of coefficients that cannot be estimated in any single study because no study obtains all of the necessary correlations. It can also be generalized to handle random effects models in which correlation parameters vary across studies. Notation and a model for the results of a series of studies examining intercorrelations among a set of "p+1" variables are provided. An estimator of the pooled correlation matrix and its standard error are derived, and a set of procedures for determining whether a common population correlation matrix underlies the set of empirical results is provided. How the pooled correlation matrix can be used to estimate the standardized coefficients of a regression model for variables measured in a series of studies is shown. Data from a synthesis of relationships among school science variables are used to illustrate the procedures. (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

BETSY JANE BECKER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED322202

Synthesizing Correlation Matrices

Betsy Jane Becker
Michigan State University

Larry V. Hedges
University of Chicago

Paper to be presented at the annual meetings of the
American Educational Research Association
Boston, MA

April 16, 1990

TM015407

BEST COPY AVAILABLE

Synthesizing Correlation Matrices

Many research syntheses which examine relationships in education and the social sciences examine only a few different bivariate relationships. Information about each relationship in the review is synthesized from a collection of primary-research studies, some of which do not examine all of the relationships of interest. The subsets of studies examining the relationships may overlap to different degrees (see, e.g., Steinkamp & Maehr, 1983), or occasionally not at all.

In some research domains, however, the parameters of investigations are more structured and may be relatively limited. Studies may look more like "replications" than in other domains. Series of studies may examine similar or identical constellations of variables, such as in the literature on the prediction of college grade-point average from entrance examination scores and high-school records. In such cases it may be desirable to combine the complete correlations matrices among the variables common to a number of studies. Methods for combining complete correlation matrices are described in this paper.

In some studies, the correlation matrices are not themselves the ultimate object of interest. In causal/explanatory studies of treatment mechanism, the correlations are of interest only because they are used to compute standardized regression coefficients corresponding to paths in the explanatory model. Meta-analysis has typically been used only to combine estimates of "first-order" effects such as treatment effects or correlation coefficients. Meta-analytic techniques developed to date have not addressed the problem of

combining information on "second-order" effects such as partial regression coefficients. Consequently meta-analysis has not been very useful in the synthesis of evidence from causal/explanatory research.

In this paper we consider the problem of combining information to estimate standardized partial regression coefficients in a linear model. A combined estimate obtained from the pooled correlation matrix is proposed and its large sample distribution is obtained. This estimate has the virtue that it can be generalized to address situations in which not every study measures every variable. In fact the method could be used to synthesize estimates of coefficients that could not be estimated in any single study because no study obtained all of the necessary correlations. It can also be generalized to handle random effects models in which correlation parameters vary across studies.

The first section below presents notation and a model for the results of a series of studies examining intercorrelations among a set of p+1 variables. Next we derive an estimator of the pooled correlation matrix and its standard error, and in the third section provide a set of procedures for determining whether a common population correlation matrix underlies the set of empirical results. Section four shows how the pooled correlation matrix can be used to estimate the standardized coefficients of a regression model for variables measured in the series of studies. Data from a synthesis of relationships among school-science variables are used to illustrate the procedures.

Notation and Model

Let $X_0, X_1, \ldots, X_p$ be random variables with the multivariate normal distribution. We consider the situation in which each of a series of k studies has examined correlations among these same $(p+1)$ variables. For example, X_0 may be an outcome and $X_1, \ldots, X_p$ may be p predictors. Each study provides $p^* = p(p+1)/2$ nonredundant correlations which are of interest. Let r_{ist} and ρ_{ist} be the sample and population correlations between X_s and X_t in the i^{th} study and let $\underset{\sim}{r}_i = (r_{i01}, r_{i02}, \ldots, r_{i0p}, r_{i12}, \ldots, r_{i(p-1)p})'$ and $\underset{\sim}{\rho}_i = (\rho_{i01}, \rho_{i02}, \ldots, r_{i0p}, \rho_{i12}, \ldots, \rho_{i(p-1)p})'$ be the vector of $p(p+1)/2$ nonredundant sample and population correlations, respectively. When it is convenient to refer to an element of the vectors $\underset{\sim}{r}_i$ and $\underset{\sim}{\rho}_i$ by sequential position, we shall do so via a Greek subscript α or γ (e.g., $r_{i\alpha}$ and $r_{i\gamma}$ are elements of $\underset{\sim}{r}_i$). Thus $\underset{\sim}{r}_i = (r_{i\alpha})$ where α runs over the range $\alpha = 1, \ldots, p(p+1)/2$.

Distribution of $\underset{\sim}{r}$. Olkin and Siotani (1976) showed the asymptotic distribution of $\sqrt{N}\,(\underset{\sim}{r}_i - \underset{\sim}{\rho}_i)$ to be normal with mean zero and variance-covariance matrix that depends on ρ_i. This implies that in large samples, $\underset{\sim}{r}_i$ is approximately normally distributed with mean vector $\underset{\sim}{\rho}_i$ and variance-covariance matrix $\underset{\sim}{\Sigma}_i$, where the elements of $\underset{\sim}{\Sigma}_i$ are defined by $\sigma_{\alpha\gamma}$,

$$\sigma_{\alpha\alpha} = \mathrm{Var}(r_{i\alpha}) = [(1 - \rho_{i\alpha}^2)]^2/n_i \quad , \quad \text{and} \qquad (1)$$

and

$$\sigma_{\alpha\gamma} = \mathrm{Cov}(r_{i\alpha}, r_{i\gamma}).$$

A formula for $\sigma_{\alpha\gamma}$ is given by Olkin and Siotani (1976, p. 238). The covariance can most easily be expressed by noting that if $r_{i\alpha} = r_{ist}$, the correlation between the s^{th} and t^{th} variables in study i; $r_{i\gamma} = r_{iuv}$, and ρ_{ist} and

ρ_{iuv} are the corresponding population values, then

$$\text{Cov}\,(r_{ist}, r_{iuv}) = [0.5\,\rho_{ist}\rho_{iuv}(\rho^2_{isu} + \rho^2_{isv} + \rho^2_{itu} + \rho^2_{itv}) + \rho_{isu}\rho_{itv} + \rho_{isv}\rho_{itu} - (\rho_{ist}\rho_{isu}\rho_{isv} + \rho_{its}\rho_{itu}\rho_{itv} + \rho_{ius}\rho_{ist}\rho_{iuv} + \rho_{ivs}\rho_{ivt}\rho_{ivu})]/n_i. \quad (2)$$

The results of k independent studies, each examining p^* correlations, can be expressed as the concatenation of the vectors $\mathbf{r}_1, \ldots, \mathbf{r}_k$ containing the nonredundant elements of the matrices of results of the k studies. Let the $kp^* \times 1$ vectors of sample and population correlations be denoted as

$$\mathbf{r} = \begin{pmatrix} \mathbf{r}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{r}_k \end{pmatrix},$$

and

$$\boldsymbol{\rho} = \begin{pmatrix} \boldsymbol{\rho}_1 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{\rho}_k \end{pmatrix},$$

respectively. If the sample sizes of the k independent studies tend to infinity at the same rate (formally if $N = \sum_{i=1}^{k} n_i$ and if the $\pi_i = n_i/N$ remain fixed as $N \to \infty$) then $\sqrt{N}\,(\mathbf{r} - \boldsymbol{\rho})$ has a nondegenerate asymptotic distribution as $N \to \infty$. This leads to the large sample approximation that $\mathbf{r}$ is normally distributed about $\boldsymbol{\rho}$. The large sample variance-covariance matrix of $\mathbf{r}$ is then $\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a blockwise diagonal matrix with submatrices $\boldsymbol{\Sigma}_1$ through $\boldsymbol{\Sigma}_k$, where $\boldsymbol{\Sigma}_i$ is defined above. Specifically,

$$\Sigma = \begin{bmatrix} \Sigma_1 & & 0 \\ & \Sigma_2 & \\ & \cdots & \\ 0 & & \Sigma_k \end{bmatrix}. \qquad (3)$$

Example. Data for the example are from two dissertations which explored the prediction of chemistry achievement for high-school students. Jensen (1966) and Smith (1966) both examined collections of aptitude, activity, and interest measures as predictors of chemistry achievement for males and females. In this example we examine correlations among measures of three variables from each study, as shown below.

Measures of Three Variables Examined

Variable	Study	
	Jensen (1966)	Smith (1966)
Chemistry Aptitude	Iowa Placement Examinations Chemistry Aptitude, Form M	Iowa Placement Examinations Chemistry Aptitude, Form M, Reading in Chemistry Subtest
Math Ability	Holzinger-Crowder Uni-Factor Tests, Numerical	Differential Aptitude Test, Numerical Ability
Verbal Ability	Holzinger-Crowder Uni-Factor Tests, Verbal	Different Aptitude Test, Verbal Reasoning

Thus each study provided two samples, and for each sample there is a 3 x 3 correlation matrix. In the ith sample, the correlations among the three variables are represented in our notation as shown below:

	Chemistry	Math	Verbal
Chemistry	1.0	r_{i1}	r_{i2}
Math	r_{i1}	1.0	r_{i3}
Verbal	r_{i2}	r_{i3}	1.0

Writing these correlations as a vector r_i, the relationships represented are

$$\begin{array}{ll} \text{Chemistry - Math} & \\ \text{Chemistry - Verbal} & \begin{bmatrix} r_{i1} \\ r_{i2} \\ r_{i3} \end{bmatrix} \\ \text{Math - Verbal} & \end{array}$$

The r_i vectors for the four samples in the example are

$$r_1 = \begin{bmatrix} .46 \\ .26 \\ .38 \end{bmatrix}, \; r_2 = \begin{bmatrix} .35 \\ .36 \\ .47 \end{bmatrix}, \; r_3 = \begin{bmatrix} .49 \\ .39 \\ .26 \end{bmatrix}, \; r_4 = \begin{bmatrix} .57 \\ .46 \\ .35 \end{bmatrix}$$

Each vector of correlations has an associated limiting variance-covariance matrix, computed using (1) and (2) above. The matrices for the two samples from Smith (1966) are

$$\Sigma_1 = \begin{bmatrix} .004 & .002 & .001 \\ .002 & .005 & .003 \\ .001 & .003 & .004 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} .007 & .004 & .002 \\ .004 & .006 & .002 \\ .002 & .002 & .005 \end{bmatrix},$$

and those for the males and females studied by Jensen are

$$\Sigma_3 = \begin{bmatrix} .0011 & .0003 & .0006 \\ .0003 & .0013 & .0009 \\ .0006 & .0009 & .0016 \end{bmatrix} \quad \text{and} \quad \Sigma_4 = \begin{bmatrix} .0010 & .0004 & .0007 \\ .0004 & .0013 & .0010 \\ .0007 & .0010 & .0016 \end{bmatrix}$$

The vector of all correlations to be synthesized then is

$r' = (.46, .26, .38, .35, .36, .47, .49, .39, .26, .57, .46, .35)$,

and its variance Σ is the 12 x 12 blockwise diagonal matrix comprised of Σ_1 through Σ_4. The sample sizes for the four samples are given below:

Sample	Sample size
Smith males	$n_1 = 174$
Smith females	$n_2 = 117$
Jensen males	$n_3 = 545$
Jensen females	$n_4 = 469$

Estimating the Pooled Correlation Matrix

When all of the studies share a common population covariance matrix, that is when $\rho_1 = \ldots = \rho_k$, it makes sense to pool estimates from the studies to estimate the common correlation matrix. Several estimates are available. Among the simpler methods is the computation of a weighted average across studies for each element of ρ. Weighting by sample size (which is proportional to the reciprocal of sampling variance) is the equivalent of carrying out a univariate meta-analysis on each component separately. Such an approach weights studies to reflect the precision of the information that they provide about a correlation. A more complex, but slightly more efficient, method of pooling is to combine via generalized least squares. This approach

weights differentially between studies but it also makes use of the information about one component (one correlation) contained in the other (correlated) components of each $\mathbf{r}_i$ vector.

Simple Weighting. Since the vector $\sqrt{N}(\mathbf{r} - \boldsymbol{\rho})$ is asymptotically normal with mean zero and variance-covariance matrix $\boldsymbol{\Sigma}$ it is straightforward to obtain the distribution of a vector of average or pooled correlations. Define the $p^* \times 1$ vector of estimated pooled correlations as

$$\mathbf{r}_\cdot = (r_{\cdot 1}, \ldots, r_{\cdot p^*})'.$$

Each element of $\mathbf{r}_\cdot$ is defined as the weighted average

$$r_{\cdot\alpha} = \sum_{i=1}^{k} w_{i\alpha} r_{i\alpha}.$$

Because the vector $\mathbf{r}_\cdot$ is a linear function of the original vector of correlations $\mathbf{r}$ the distribution of $\mathbf{r}_\cdot$ is obtained via the multivariate δ method (e.g., Rao, 1972). The multivariate δ method states that if $\sqrt{N}(\mathbf{r} - \boldsymbol{\rho})$ has a multivariate normal distribution with a mean of zero and variance-covariance matrix $\boldsymbol{\Sigma}$, then $\sqrt{N}(f(\mathbf{r}) - f(\boldsymbol{\rho}))$ is normal with variance-covariance matrix $\mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}'$ where $\mathbf{A}$ is the $p^* \times kp^*$ matrix of first derivatives of $f(\mathbf{r}) = \mathbf{r}_\cdot$ with respect to $\mathbf{r}$, evaluated at the value $\boldsymbol{\rho}$ (i.e., $\mathbf{A}$ is the Jacobian matrix).

In the present example the vector of pooled correlations is $\mathbf{r}_\cdot = f(\mathbf{r})$ as defined above. Therefore the matrix of partial derivatives $\mathbf{A}$ equals

$$\mathbf{A} = \begin{bmatrix} \underline{w}_{11} & 0\ldots & 0 & \underline{w}_{21} & 0\ldots & 0\ldots & \underline{w}_{k1} & 0\ldots & 0 \\ 0 & \underline{w}_{12} & 0\ldots & 0 & \underline{w}_{22} & 0\ldots & 0 & \underline{w}_{k2} & \ldots 0 \\ & & \cdot & & & \cdot & & \cdot & \\ & & \cdot & & & \cdot & & \cdot & \\ & & \cdot & & & \cdot & & \cdot & \\ 0\ldots & & \underline{w}_{1p^*} & 0\ldots & & \underline{w}_{2p^*} & \ldots & 0\ldots & \underline{w}_{kp^*} \end{bmatrix} .$$

Applying the multivariate δ method produces the variance-covariance matrix $\mathbf{V} = \mathbf{A}\,\mathbf{\Sigma}\,\mathbf{A}'$ with elements $\underline{V}_{\alpha\gamma}$, such that

$$\underline{V}_{\alpha\gamma} = \sum_{j=1}^{k} \underline{w}_{j\alpha}\,\underline{w}_{j\gamma}\,\sigma_{j\alpha\gamma} , \tag{4}$$

where $\sigma_{j\alpha\gamma}$ is given by (2) above. The diagonal elements of V (the variances) simplify to weighted sums of the variances of the individual $\underline{r}$ values, specifically,

$$V_{\alpha\alpha} = \sum_{j=1}^{k} \underline{w}_{j\alpha}^2\,\sigma_{j\alpha}^2 ,$$

where $\sigma_{j\alpha}^2$ is defined in (1).

The weight values. The results above hold for a variety of values of $\underline{w}_{i\alpha}$. Two obvious possibilities are weighting the correlations according to sample size and by their precision. Specifically, we might simply weight each $\underline{r}_{i\alpha}$ by its sample size via

$$\underline{r}_{\cdot\alpha} = \frac{\sum_{j=1}^{k} n_j\,\underline{r}_{j\alpha}}{\sum_{j=1}^{k} n_j} ,$$

which is equivalent to using the weight

$$w_{i\alpha} = n_i \Big/ \sum_{j=1}^{k} n_j \qquad (5)$$

Because the large sample variances of $r_{i\alpha}$ is $(1-\rho_{i\alpha}^2)^2/n_i$, weighting by n_i is equivalent to weighting by the precision (i.e., the inverse of the variance) of each correlation.

A generalized least squares approach. A slightly more complex estimate can also be obtained by considering the vector **r** as outcome data for a generalized least squares regression analysis. Specifically to estimate a common correlation vector of length p^*, the model is

$$\mathbf{r} = \mathbf{X}\underline{\rho} + \mathbf{e},$$

where **r** is defined as above, $\underline{\rho}' = (\rho_1, \rho_2, \ldots, \rho_{p^*})$ is the set of common correlations to be estimated, and **X** is a matrix of dimension $kp_* \times p_*$ created by "stacking" of k identity matrices, each of dimension $p^* \times p^*$. For the data in our example, **X** is the 12 x 3 matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

In the generalized least squares model, the error vector has approximate covariance matrix Σ given in (3) so that the generalized least squares estimate of $\underline{\ell}$ is given by

$$(\mathbf{X}' \, \Sigma^{-1} \, \mathbf{X})^{-1} \, \mathbf{X}' \, \Sigma^{-1} \, \mathbf{r} \quad .$$

The approximate variance covariance matrix for this estimate is given by

$$(\mathbf{X}' \, \Sigma^{-1} \, \mathbf{X})^{-1} \quad .$$

Example. Weighting by the inverse of the variance of each correlation (i.e., w_{ik} is defined in (5)) produces a vector of pooled correlations $\mathbf{r}.$, where

$$\mathbf{r}' = (.51, \ .40, \ .33).$$

The variance-covariance matrix V, with elements defined by (4), is

$$\mathbf{V} = \begin{bmatrix} .0004 & .0002 & .0002 \\ .0002 & .0005 & .0004 \\ .0002 & .0004 & .0006 \end{bmatrix} .$$

Testing the Hypothesis that the Correlation Matrices Are Homogeneous

In the previous section, we described a method for estimating a common population correlation matrix from several studies. Summarizing the data from several studies by a single estimated correlation matrix is conceptually sensible only if the studies have the same population correlation matrix. To determine whether the data obtained from several studies is reasonably consistent with the hypothesis of a common correlation matrix, it is useful to have a formal hypothesis test. In this section we describe a test for

homogeneity of correlation matrices across studies. It is a slight generalization of the tests reported in Chapter 10 of Hedges and Olkin (1985).

Let $\underset{\sim}{\rho}_1, \ldots, \underset{\sim}{\rho}_k$ and $\mathbf{r}_1, \ldots, \mathbf{r}_k$ be the vectors of length p(p+1)/2 including the nonredundant population and sample correlations from each of k studies, and let $\mathbf{V}_1, \ldots, \mathbf{V}_k$ be the large sample covariance matrices of $\mathbf{r}_1, \ldots, \mathbf{r}_k$ given in the previous section. Define the k(p+1)/2 dimensional vector r, the kp(p+1)/2 x p(p+1)/2 matrix **X**, and the kp(p+1)/2 x kp(p+1)/2 matrix V by

$$\mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{r}_k \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{I}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{I}_k \end{bmatrix}$$

$$\mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_k),$$

where $\mathbf{I}_1, \ldots, \mathbf{I}_k$ are identity matrices of order p(p+1)/2.

A test of the hypothesis of homogeneity of correlation matrices across studies, that is to test

$$H_0 : \underset{\sim}{\rho}_1 = \ldots = \underset{\sim}{\rho}_k,$$

uses the statistic

$$Q = \mathbf{r}' \left[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\right]\mathbf{r}. \qquad (6)$$

When H_0 is true then Q has approximately a chi-squared distribution with (k-1)(p+1)p/2 degrees of freedom. Thus a test of H_0 at the 100α percent level of significance is given by rejecting H_0 if Q exceeds $100(1-\alpha)$ percentile point of the chi-squared distribution with (k-1)(p+1)p/2 degrees of freedom.

Example. Returning to our example and computing Q via (6) we obtain Q = 97.88. Comparing 97.88 with 34.81 the 99 percent critical value of the chi-squared distribution with (4-1)(3+1)3/2 = 18 degrees of freedom, we see that

the hypothesis of homogeneity of correlation matrices is rejected at the 1% level of significance. (We will, however, continue with this example for the purposes of illustration.)

Estimating a Linear Model from a Pooled Correlation Matrix.

In this section we consider estimation of standardized regression coefficients from a pooled correlation matrix. We begin by deriving the asymptotic joint distribution of the standardized regression coefficients from an arbitrary consistent, asymptotically normally distributed estimate of the correlation matrix. Using the pooled estimate of the correlation matrix and the asymptotic distribution obtained in the last section, the results given here yield the asymptotic distribution of the pooled estimate of the standardized regression coefficients. Note that when $k = 1$, these results also yield the asymptotic distribution of the standardized regression coefficents computed from a single sample.

Background. Let $X_0, X_1, \ldots, X_p$ be random variables with a multivariate normal distribution and population correlation matrix P. Partitioning P as given below

$$P = \begin{pmatrix} 1 & P_{10} \\ P_{01} & 1 \end{pmatrix}$$

the (population) standardized regression coefficients $\beta = (\beta_1, \ldots, \beta_p)'$ of the regression of X_0 on $X_1, \ldots, X_p$ are given by

$$\beta = P_{11}^{-1} P_{01} .$$

If R is a sample estimate of P and R is partitioned as

$$R = \begin{bmatrix} 1 & R_{12} \\ R_{21} & R_{11} \end{bmatrix}$$

a sample estimate $b = (b_1, \ldots, b_p)'$ of β is given by

$$b = R_{11}^{-1} R_{21} . \tag{7}$$

Notation. Let $r_{.} = (r_{.21}, r_{.32}, \ldots r_{.p1}, r_{.13}, \ldots, r_{.(p-1)p})'$ and $\rho = (\rho_{21}, \rho_{32}, \ldots \rho_{p1}, \rho_{13}, \ldots \rho_{(p-1)p})'$ denote the vector of $p(p+1)/2$ nonredundant sample estimates and population correlations respectively. When it is convenient to refer to an element of the vectors $r_{.}$ or ρ by sequential position, we shall do so via a Greek subscript α or γ (e.g., $r_{.\alpha}$ and $r_{.\gamma}$ are elements of $r_{.}$). Thus $r_{.} = (r_{\alpha})$, where α runs over the range $\alpha = 1, \ldots, (p+1)p/2$.

Result. If R is a sample estimate of the correlation matrix P with an asymptotic distribution as $N \to \infty$ given by

$$\sqrt{N}\,(r_{.} - \rho) \sim N(0, (\rho))$$

where $\Psi(\rho)$ is a function of ρ, then the asymptotic distribution of the standardized regression coefficient estimates b is given by

$$\sqrt{N}\,(b - \beta) \sim N(0, \Phi(\rho))$$

where $\Phi(\rho) = A\,\Psi(\rho)\,A'$, $A(\rho) = (a_{\alpha\beta})$,

$$a_{\alpha\beta} = \begin{cases} -\sum_{j=1}^{p} (\rho^{\alpha j}\rho^{\beta j} + \rho^{\alpha j}\rho^{\beta j})\,\rho_{j\beta} & \text{if } \rho_{\alpha} = \rho_{\alpha j} \in P_{11} \\ \rho^{\alpha\beta} & \text{if } \rho_{\alpha} = \rho_{\alpha\beta} \in P_{21}, \end{cases} \tag{8}$$

and $\rho^{\alpha\beta}$ are the elements of P^{-1} .

This result implies that the large samples multivariate normal approximation to the distribution of b is

$$b \sim N(\beta, S(r))$$

where $S(r) = A(r)\,\Psi(r)\,A'(r)/N = A(r)\,V\,A'(r)$, $A(r)$ and $\Psi(r)$ are defined by replacing each population correlation in A and Ψ by the corresponding sample correlation, and V is the large sample covariance matrix of r..

Proof. Given that $\sqrt{N}\,(r - \rho)$ is asymptotically normal, use the multivariate delta method to obtain the asymptotic distribution of $b = f(r) = [f_1(r), \ldots, f_p(r)]$. This gives the asymptotic distribution of $\sqrt{N}\,[f(r) - f(\rho)]$ as $N(0, A\,\Psi\,A')$ where $A = (a_{i\alpha})$ and

$$a_{i\alpha} = \left.\frac{\partial\, f_i(r)}{\partial\, r_\alpha}\right|_{r = \rho}.$$

Partitioning P and R as indicated above,

$$b = R_{11}^{-1}\,R_{21}.$$

Using the fact that

$$\frac{\partial b}{\partial r_\alpha} = \frac{\partial R_{11}^{-1}}{\partial r_\alpha}\,R_{21} + R_{11}^{-1}\,\frac{\partial R_{21}}{\partial r_\alpha}$$

and

$$\frac{\partial\, R^{-1}}{\partial r_{ij}} = -R^{-1}\,(J_{ij} + J_{ji})\,R^{-1}$$

(where J_{ij} is a matrix with a one in the i^{th} row and j^{th} column and zeroes elsewhere) yields the result.

Application. The result given above can be used to establish confidence intervals for the standardized regression coefficients β_i or to test their statistical significance. A 100(1-α) percent confidence interval for β_i is given by

$$b_i - z_{\alpha/2}\,\sqrt{s_{ii}}\,/\,\sqrt{N} \le \beta_i \le b_i + z_{\alpha/2}\,\sqrt{s_{ii}}\,/\,\sqrt{N} \qquad (10)$$

where s_{ii} is the i^{th} diagonal element of S(r) and z_α is the 100α percent critical value of the standard normal distribution.

The corresponding test of the statistical significance of β_i (that is, a test of the hypothesis that $\beta_i = 0$) uses the statistic

$$T = \sqrt{N}\, b_i \,/\, \sqrt{s_{ii}}.$$

If $|T_i| > z_{\alpha/2}$, the hypothesis that $\beta_i = 0$ is rejected with (two-tailed) significance level α.

The trivariate normal distribution. The trivariate normal distribution is the simplest nontrivial case that illustrates the computations. In this case the two predictor variables are X_1 and X_2 and X_0 is the criterion. The full correlation matrix is

$$R = \begin{pmatrix} 1 & r_{01} & r_{02} \\ r_{01} & 1 & r_{12} \\ r_{02} & r_{12} & 1 \end{pmatrix} = \begin{pmatrix} 1 & R_{10} \\ R_{01} & R_{11} \end{pmatrix},$$

where

$$R_{11} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \qquad R_{01} = \begin{pmatrix} r_{01} \\ r_{02} \end{pmatrix}.$$

The estimated standardized regression coefficients are

$$b_1 = \frac{r_{10} - r_{20}\, r_{12}}{1 - r_{12}^2}, \qquad b_2 = \frac{r_{20} - r_{10}\, r_{12}}{1 - r_{12}^2}.$$

In vector form the correlation matrix is $\mathbf{r}_{\cdot} = (r_{01}, r_{02}, r_{12})'$.

Using the previous result we obtain the matrix $\mathbf{A}(\mathbf{r})$ as

$$\mathbf{A} = \begin{pmatrix} \dfrac{1}{1-r_{12}^2}, & \dfrac{-r_{12}}{1-r_{12}^2}, & \dfrac{2\, r_{10}\, r_{12} - r_{20} - r_{12}^2\, r_{20}}{(1-r_{12}^2)^2} \\[2ex] \dfrac{-r_{12}}{1-r_{12}^2}, & \dfrac{1}{1-r_{12}^2}, & \dfrac{2\, r_{20}\, r_{12} - r_{10} - r_{12}^2\, r_{10}}{(1-r_{12}^2)^2} \end{pmatrix}.$$

Hence the vector of standardized regression coefficients $b = (b_1, b_2)'$ has approximately a bivariate normal distribution given by

$$b \sim N(\beta, A V A'),$$

where $\beta = (\beta_1, \beta_2)'$ are the population standardized regression coefficients

$$\beta_1 = \frac{\rho_{10} - \rho_{20}\rho_{12}}{1 - \rho_{12}^2}, \qquad \beta_2 = \frac{\rho_{20} - \rho_{10}\rho_{12}}{1 - \rho_{12}^2}$$

and V is the approximate convariance matrix of r. given in the previous section.

Example. Return to our example and consider the predictability of chemistry aptitude (X_0) from mathematical ability X_1 and verbal ability X_2. In this case

$$R_{11} = \begin{bmatrix} 1.0 & .33 \\ .33 & 1.0 \end{bmatrix}, \qquad R_{01} = \begin{bmatrix} .51 \\ .40 \end{bmatrix},$$

so that $b = R_{11}^{-1} R_{01} = (.42, .26)'$. The matrix A of partial derivatives computed using (8), but substituting sample estimates for population correlations, is

$$A = \begin{bmatrix} 1.126 & -.376 & -.228 \\ -.376 & 1.126 & -.226 \end{bmatrix}.$$

Consequently, the variance-covariance matrix of b is

$$A V A' = 10^{-4} \begin{bmatrix} 4.4 & -2.5 \\ -2.5 & 4.94 \end{bmatrix}.$$

Thus the standard error of b_1 is .021, the standard error of b_2 is .022, and

the correlation between the estimates b_1 and b_2 is -.54. Ninety-five percent confidence intervals for β_1 and β_2 using (10) are

$$.38 = .42 - 1.96(.021) \leq \beta_1 \leq .42 + 1.96(.021) = .46,$$

and

$$.22 = .26 - 1.96(.022) \leq \beta_2 \leq .26 + 1.96(.022) = .30.$$

Since neither confidence interval contains zero, both standardized regression coefficients are significantly greater than zero.

References

Hedges, L.V. & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Academic Press.

Jensen, J.A. (1966). An analysis by class size and sex of orthogonalized interest and aptitude predictors in relation to high school chemistry achievement and criteria. Unpublished doctoral dissertation, University of Rochester, N.Y.

Olkin, I. & Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. In S. Ideka (Ed.) Essays in probability and statistics. Tokyo: Sinko Tsusho.

Rao, C.R. (1971). Linear statistical inference. New York: Wiley.

Smith, I.R. (1966). Factors in chemistry achievement among eleventh-grade girls and boys. Unpublished doctoral dissertation, The Catholic University of America.

Steinkamp, M.W. & Maehr, M.L. (1983). Affect, ability and science achievement: A quantitative synthesis of correlational research. Review of Educational Research, 53, 369-396.